基于决策树的敏感词变形体识别算法研究及应用 *

余敦辉 1, 2，张笑笑 1†，付 聪 1，张万山 1, 2

(1. 湖北大学 计算机与信息工程学院，武汉 430062; 2. 湖北省教育信息化工程技术中心，武汉 430062)

摘 要：针对网络中敏感词变形体识别效率不高的问题，提出了基于决策树的敏感词变形体识别算法。首先，通过分析汉字的结构和读音等特征，研究敏感词及变形体；其次，基于敏感词库构建敏感词决策树；最后，通过多因子改进模型，对微博等新媒体的文本敏感程度进行计算。实验结果表明，该算法在识别中文敏感词及变形体时，查全率和查准率最高分别可达 95%和 94%，与基于确定有穷自动机的改进算法相比，查全率和查准率分别提高 19.8%和 21.1%；与敏感信息决策树信息过滤算法相比，查全率和查准率分别提高 17.9%和 18.1% 。通过分析，该算法对敏感词变形体的识别和自动过滤是有效的。

关键词：敏感词识别；敏感词变形体；决策树；敏感程度计算；多因子模型

中图分类号：TP391.1 doi: 10.19734/j.issn.1001-3695.2018.11.0792

Research and application of change form of sensitive words recognition algorithm based on decision tree

Yu Dunhui[1, 2], Zhang Xiaoxiao[1†], Fu Cong[1], Zhang Wanshan[1, 2]

(1. *College of Computer & Information Engineering, Hubei University, Wuhan* 430062, *China*; 2. *Education Informationization Engineering & Technology Center of Hubei Province, Wuhan* 430062, *China*)

Abstract: In order to solve the problem that the recognition efficiency of sensitive word deformed bodies of the network text is not high, this paper proposed a sensitive word deformed bodies recognition algorithm based on decision tree. Firstly, it studied sensitive words and deformed bodies by analyzing the characteristics of Chinese characters and pronunciation and so on. Secondly, it constructed a sensitive word decision tree based on sensitive word library. Finally, it calculated the text sensitivity of new media such as Weibo by multi-factor improvement model. The experimental results show that the proposed algorithm can achieve the highest recall rate and precision rate of 95% and 94% respectively when identifying Chinese sensitive words and deformed bodies. Compared with the improved algorithm based on the finite automaton, the recall rate and the precision rate are increased by 19.8% and 21.1% respectively. Compared with the sensitive information decision tree information filtering algorithm, the recall rate and the precision rate are increased by 17.9% and 18.1%. The analysis show that the algorithm is effective in the recognition and automatic filtering of sensitive word deformed bodies.

Key words: sensitive word recognition; sensitive word deformable body; decision tree; sensitivity computation; multi factor model

0 引言

随着互联网的快速发展，网络信息呈指数级增长，非法言论（如黄赌毒、恐怖、暴力血腥信息）经常充斥其中[1,2]，这些不良信息通常带有一些敏感词汇，并大量以变形体的形式出现，给民众尤其是青少年带来了巨大的伤害，对国家安全、社会稳定和网络环境的健康形成严重威胁。微博作为新型的广播式社交网络平台，以实时、便捷的特点广泛传播、分享和获取简短信息，但由于用户群庞大及监管能力有限等因素，不法分子经常将敏感词汇散布其中。因此，对于微博等新媒体中的敏感词及其变形体的识别和过滤已经成为了迫切需要解决的研究课题。

目前众多学者对敏感词及其变形体的识别和过滤问题纷纷展开研究。文献[3]提出了通过构建 CNN-like 词网对文本敏感词进行分析处理，提高了敏感词检测的准确率，缺点是需要对人工构建词与词之间的关联。文献[4]提出一种基于确定有穷自动机的改进算法，通过敏感词拼音的第一个字母来构建敏感信息决策树，其优点是不依赖敏感信息语料库，能够提高敏感信息的检测效率；缺点是对敏感词变形体无处理能力。文献[5]提出一种敏感信息决策树信息过滤算法，同样通过构建敏感词决策树提高检索速度，并通过给出敏感词权重以达到敏感文本检测和过滤的目的。该方法依赖人工确定敏感级别，难以客观地表现文本的敏感程度，而且缺少对敏感词变形体的分析和识别。文献[6]针对变异的敏感词汇提出了一种方法，该方法将某些特殊字符转换成形状相似的字母，然后再进行检测，但是对变异的变形体识别效率不高。文献[7]采用机器学习的方法，通过采用 bigram、词干等作为特征值来对文本信息做分类分析，以检测出变形体。这些方法对英文字符有较好的处理效果，但没有将中文敏感词变形体考虑在内。

收稿日期：2018-11-20；**修回日期：**2019-01-16 **基金项目：**国家重点研发计划资助项目（2016YFB0800401）；国家自然科学基金资助项目（61572371，61832014）；湖北省技术创新专项（重大项目）（2018ACA13）

作者简介：余敦辉（1974-），男，湖北武人，副教授，博士，主要研究方向为服务计算、大数据；张笑笑（1995-），女（通信作者），山东滨州人，硕士研究生，主要研究方向为大数据，舆情监测（201811111911402@stu.hubu.edu.cn）；付聪（1991-），男，湖北武人，硕士研究生，主要研究方向为大数据；张万山（1973-），男，湖北武人，讲师，硕士，主要研究方向为 Web 信息挖掘.

总之，目前在中文敏感词变形体识别与过滤的研究中，存在对敏感词变形体分析不足、识别与过滤效率偏低等问题。为此，本文提出了基于决策树的敏感词变形体识别算法（recognition of sensitive words based on decision tree, RSWDT），着力解决敏感词变形体识别与过滤问题。首先，根据汉字的读音和结构，分析词的拼音模式、词的简称和词的拆分三种敏感词变形体模式；然后，扩充现有的敏感词库，在敏感词库中增加词的拼音、区位码以及拆分后的区位码等信息，进而根据敏感词库构建敏感词决策树来实现对敏感词变形体的准确识别；最后，结合改进的多因子模型，针对微博、博客、网络评论等网络文本，计算文本敏感程度，实现敏感文本的自动过滤。

1　问题描述

1.1　敏感词变形体分析及处理

本文研究的敏感词变形体包括词的拼音模式、词的简称模式和词的拆分模式[8]。由于现在各个网络平台对信息的审查越来越严格，很多网络文本中的敏感词以变形体形式出现，包括词的拼音模式、词的简称模式和词的拆分模式。以词“贩卖毒品”为例，其变形体的具体结构如图 1 所示。

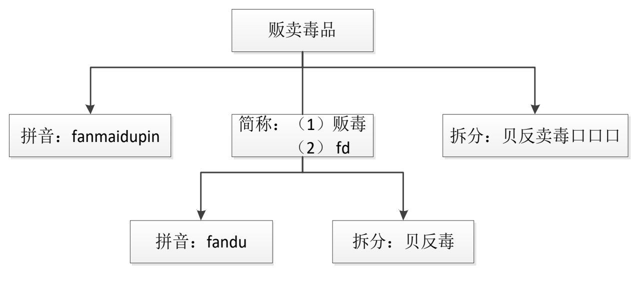


图 1　词变形体结构

Fig. 1　Structure diagram of change form of word

1）词的拼音模式

在中文中，同一个汉语拼音可能对应多个不同的汉字。如文献[8]所示，本文采用三位音码的形式，即将汉字的拼音表示为声母/韵母和声调三部分，并将每部分分别采用英文字母 a，b，……，等来进行编码，从而将一个汉字转换成一系列的字符序列即音码，将以便进行下一步的计算和比较。

2）词的简称模式

据统计，在汉语新闻文章里，在 20%左右的句子可能含有缩略语[9]，包括首字母缩写和词的缩写。其中首字母缩写如“法轮功”缩写为“flg”。词的缩写的一般分为压缩、节略和统括[10] 三种形式，其中又以压缩和节略的组合生成模式最为常见。压缩是指把全称分割为几个词语，然后从每个词语中抽取最能代表原义的汉字保留，如“贩卖\毒品”的简称为“贩毒”;节略是指在全称中直接省去部分词语，留下另一部分词语作为简称，如“复旦\大学”的简称为“复旦”。压缩和节略的基本思想都是从全称中选取部分汉字或者词语重组形成简称。在重组的过程中，字序一般不会发生改变。简称中的汉字全部包含于词的全称中，因此，找到词全称的子集就可以找到其简称。

3）词的拆分模式

根据汉字的构成单位可把汉字分为独体字、合体字两类。独体字（日、月等）由笔画构成，合体字（休、取等）则由偏旁构成。汉字的空间上的关系有相交、相离、相接[11]。汉字的方位上的关系有上下、左右，内外、框架、独体。区位码是一个四位的十进制数，每个区位码都对应着一个唯一的汉字或符号。根据以上汉字特征对敏感词列表中的汉字进行人工拆分，并采用区位码进行编码形成汉字拆分表，如表 1 所示。

表 1　汉字拆分表

Table 1　Table of word splitting

汉字	区位码	拆分	区位码
法	2308	氵 去	6763 4005
秃	4526	禾几	2644 2824
国	2590	口玉	3158 5181

为了识别出敏感词拆分的变形体，首先根据汉字拆分表把敏感词与疑似敏感词变形体进行拆分，并转换成相应的区位码。

1.2　整体方案

为达到敏感信息自动过滤的目的，采取以下步骤：

a）基于决策树的敏感词变形体识别。

对词的拼音、词的简称和词的拆分三种敏感词变形体，根据查找敏感词库、构建决策树、利用决策树对敏感词识别的方法，最后提出了基于决策树的敏感词变形体识别算法。

b）基于多因子改进模型的敏感信息自动过滤。

根据识别出的敏感词及其变形体，考虑文本中敏感词的位置、频繁度以及类别等因素，并基于改进后的多因子模型，对文本的敏感程度进行计算，然后根据文本的敏感程度大小对文本进行相应的处理，进而达到自动过滤的效果。

2　基于决策树的敏感词变形体识别

2.1　敏感词决策树建立

通过对敏感词及其变形体的分析，了解到对敏感词变形体的识别——需对每个汉字做拼音、音码以及区位码分析。因此，为了准确地查询和匹配敏感词变形体，需在决策树构建之前，对已有的敏感词库（表 2）进行信息扩充，用于存储已知敏感词汉字的拼音、音码以及区位码等相关信息，便于决策树的建立及存储，为建立决策树提供依据。

表 2　敏感词库

Table 2　Library of sensitive words

首字母	首汉字及拼音	音码	区位码	拆分后区位码	敏感词
A	安（an）	HA	1618	6918 3714	安眠药，安乐死……
B	八（ba）	AAA	1643	1643	八嘎……
C	草（cao）	TFC	1861	6019 5271	草泥马……
……	……	……	……	……	……
Z	作(zuo)	SXD	5587	5674 5307	作死……

决策树构建算法将敏感词库中的敏感词按第一个字的拼音首字母分类。同时，再对首字母进行同字聚类，使首字母相同的敏感词在一个分支下，使相同的字只存储一次，便于提高检索速度，节约存储空间。在节点存储汉字的同时，将该汉字的拼音、音码及对应的区位码也存储起来。叶子节点用于记录算法识别出的敏感词或者敏感词变形体的位置以及类别。其中每个叶子节点敏感词位置与类别信息的下标，根据类哈夫曼编码规则制定，当分支数大于 2 时，用实际分支数标记。当出现词的拼音、词的简称和词的拆分时，决策树识别算法也同样能够将其检测出来，如“安 mian 药”“亻乍死”。

将敏感词库作为决策树构建算法的输入，输出一棵敏感词决策树，如图 2 所示。

敏感词库 $S=\{s_0,s_1,\ldots,s_i,\ldots,s_{n-1}\}$，$(0<i<n)$，n 为敏感词个数，$s_i$ 表示敏感词；$s_i=\{s_{i,0},s_{i,1},\ldots,s_{i,j},\ldots,s_{i,l-1}\}$，$(0<j<l)$，$s_{ij}$ 表示第 i 个敏感词的第 j 个敏感字，l 表示敏感词长度。

为识别敏感词变形体，通过敏感词决策树构建算法

（establishment of sensitive word decision tree algorithm，ESDT），首先将敏感词库作为输入，建立根节点；然后通过首字母建立分支。若敏感词库的敏感词汉字信息与决策树节点信息匹配时，查找其孩子节点；若匹配继续向下查找，不匹配时建立新节点，并存储敏感词汉字及其拼音、音码及区位码；若不匹配，查找兄弟节点是否匹配，若匹配建立兄弟节点的孩子节点，并存储敏感词汉字信息；否则建立新的节点，直至将敏感词中的所有汉字存储完毕，再建立叶子节点，用于存储敏感词的位置和类别信息。然后输入下一个敏感词，直至敏感词库的敏感词全部输入建立完毕。最后输出敏感词决策树。

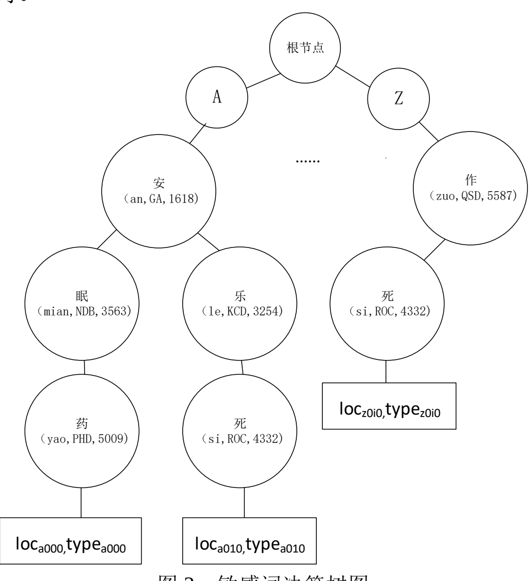


图 2　敏感词决策树图

Fig. 2 Sensitive word decision tree graph

算法具体执行过程如算法 1 所示。

算法 1　ESDT algorithm

输入：敏感词库 S 。

输出：敏感词决策树。

a) 初始化 $i=0$， $j=0$， $k=0$， k 记录孩子节点序号。

b) 输入敏感词 s_i，获取其中文长度为 l，并获取其首字母 I 。

c) 进入 I 子树查询，将 s_{ij} 与 I 的第 k 个孩子节点 $child_k$ 比较。

d) 若 $s_{ij}==child_k$ 节点的值，则 $j++$， $j<l$， $s=child_k$， $k=0$，返回步骤 c)； $j\geq l$ ，$i++$， $i<n$，返回步骤 b)；否则，结束。

e) 若 $s_{ij}\neq child_k$，查找 $child_k$ 的兄弟节点是否为空。

6.若 $child_k$ 兄弟节点为空，创建新节点 $child_{k+1}$，值为 s_{ij}，记录 s_{ij} 的拼音、音码以及区位码， $j++$。

f) 若 $j<l$，创建子节点，并赋值 s_{ij}，记录 s_{ij} 的拼音， $j++$；转步骤 f)继续处理；否则，最后一个节点记录敏感词相关信息，初始化敏感词的位置及类别， $i++$； $i<n$，返回步骤 b)；否则，结束。

g) 否则，若 $child_k$ 兄弟节点不为空， $k++$，返回步骤 b)，处理下一个敏感词；

h) 算法结束。

本文算法构建的敏感词决策树深度为敏感词库中最长敏感词的长度，一般<10。树中每个节点都存储了敏感词汉字以及其对应的拼音、音码以及区位码，叶节点还记录了敏感词的位置和类别信息，并将各个词的位置和类别都进行了初始化。

2.2　敏感词变形体识别

为了准确识别出敏感词及其变形体，算法首先获取含有疑似敏感词变形体的文本；然后输入敏感词决策树中，获取首字母进入分支，将疑似敏感词的字符与决策树中的节点信息进行匹配。当疑似敏感词变形体为汉字与拼音时，此为词的拼音或者词的简称模式，直接进行匹配；当疑似敏感词为特殊字符或者汉字部首时，获取区位码，在拆分表中进行匹配。若相同进行下一个字符比较，不同则查找孩子节点及兄弟节点，直至叶子节点。若匹配成功，在叶子节点记录敏感词及变形体位置及类别。然后输入下一个疑似敏感词，继续匹配；最后输出敏感词决策树叶子节点的信息，即敏感词及其变形体对应位置及类别信息。算法具体的执行过程如算法 2 所示。

算法 2　RSWDT algorithm

输入：敏感词决策树，含疑似敏感词变形体文本 T ， $T=\{t_0,t_1,\ldots,t_i,\ldots,t_{n-1}\}$，$(0<i<n)$， t_i 为文本字符，n 为文本字符个数。

输出：决策树叶子节点信息，敏感词及其变形体。

a) 初始化 $i=0$， $k=0$， k 用于记录第一个进入分支的字符序列号。

b) 输入文本字符 t_i， $j=0$， $k=i$ ，判断 t_i 是否为中文、英文字符或者汉字偏旁部首。如果是中文字符，需要提取首字母 I ，英文直接获取，汉字偏旁部首获取其区位码。

c) 将 t_i 与 I 的孩子 $child_j$ 相匹配。

d) 若 $t_i==child_j$ 节点值， $i++$ 。若 $i\leq n$，将 t_i 与 $child_j->child$ 进行匹配；否则 $i>n$，算法结束。

e) 若 $child_j->child=NULL$，在叶子节点记录位置 loc_l 、 $Type_l$，输出敏感词或者敏感词变形体转到步骤 b)处理。

f) 若 $child_j->child\neq NULL$ ，转到步骤 d)处理。

g) 若 $t_i\neq child_j$ 的值，将 t_i 与 $child_j->child$ 匹配，直到 $child_j->child=NULL$ 查询 $child_j$ 兄弟节点是否为空。

h) 若兄弟节点不为空，则 $j++$，转步骤 c)；

i) 若兄弟节点为空，则 $i=k+1$，若 $i\leq n$，则转到步骤 b)处理；若 $i>n$，则算法结束。

j) 算法结束。

3　基于多因子模型的敏感信息过滤

多因子模型常被应用于金融领域，用于量化投资时，综合所选取的多个因子，针对投资决策进行最终的判断。在此，本文选取文本中敏感词所处的位置、文本中敏感词所属的类别和文本中敏感词出现的频繁度等作为文本敏感信度计算因子，对包含敏感信息的文本进行自动过滤。为此本文通过决策树输出叶子节点信息中的敏感词及其变形体位置和类别，计算出文本敏感程度，步骤如下：

a)输出叶子节点中每个敏感词的位置信息，构成位置信息集合。

b)计算每个敏感词或者敏感词变形体的位置敏感度，其中根据位置信息次数的累加，可得到频繁度信息，即位置信息计算已包含对频繁度计算。

c 再根据敏感词类型表，查找每个敏感词和敏感词变形体所属类型。每个类型占有不同的权重，将其与位置敏感度进行计算，得到每个词的敏感度。

d)将每个词的敏感度累计，得到文本的敏感程度。文本的敏感程度可以辅助文本自动审查的完成。

3.1　词的拼音模式敏感词的位置信息获取

从决策树中获取全部敏感词及变形体集合 $S=\{s_0,s_1,\ldots,s_i,\ldots,s_{n-1}\}$ $(0<i<n)$ 及其每个敏感词的位置信息，构成敏感词位置信息集合 $Loc=\{l_0,l_1,\ldots,l_i,\ldots,l_{n-1}\}$ $(0<i<n)$，其中：n 为敏感词个数； l_i 表示敏感词在文本中的所在位置，用于计算位置敏感度。

3.2　敏感词及变形体位置敏感度计算

由于信息太多，为了在最短的时间内获取到更多的信息，人们往往只对信息的头部与尾部进行浏览，这也符合人们总是喜欢把概括性描述写在文章的头部与尾部的习惯。因此，敏感词出现在文本的头部对文本敏感程度的影响要比敏感词出现在尾部对文本敏感程度的影响要大，敏感词出现在文本的尾部对文本敏感程度的影响要比敏感词出现在文本其他位置对文本敏感程度的影响要大。敏感词 s_i 的位置敏感度如下：

$$s_{loc}(s_i)=\begin{cases}\alpha & 0<l_i\le a\\ \beta & a<l_i\le b\\ \lambda & b<l_i\le len(t)\end{cases} \quad (1)$$

其中：α、β、λ 表示敏感词 s_i 分别位于文本头部、中部、尾部的位置权重；$a=\frac{len(t)}{3}$，a 为文本 t 头部与中部的分界值，$b=\frac{len(t)\times 2}{3}$，b 为文本 t 中部与尾部的分界值；$l_i$ 为敏感词 s_i 对应的位置信息。

3.3　敏感词及变形体类别敏感度计算

本文依据新华社发表的禁用词规定，可将敏感词分为时政社会生活类、法律法规类、民族宗教类、港澳台和领土主权类、国际关系类五大类。表 3 是每个类别的敏感词部分例子。

表 3　敏感词分类举例

Table 3　Examples of classification of sensitive words

敏感词类别	敏感词举例
时政社会生活类	装逼，草泥马，特么的
法律法规类	罪犯，工人小偷，检查院院长
民族宗教类	鲜族，回回，蛮子
港澳台和领土主权类	内港，内澳，台独
国际关系类	北朝鲜，穆斯林国家，阿拉伯民兵

每个类别中的敏感词对文本敏感程度的影响是不同的，为此需要确定这五个类别之间的相对权重，本文采用层次分析法[12,13]给出了一种权值的计算方法。

根据上一节敏感词的分类，本文假设敏感词的类别集合为 $T=\{T_1,T_2,T_3,T_4,T_5\}$，其中 T_1 表示时政社会生活类；T_2 表示法律法规类；T_3 表示民族宗教类；T_4 表示港澳台和领土主权类；T_5 表示国际关系类。本文采用层次分析法确定敏感词类别的相对权重，得到 $typ(s_i)$。$typ(s_i)$ 表示敏感词 s_i 所对应类别的相对权重。

3.4　文本敏感度计算

以多因子模型选取出来的敏感词所处的位置、所属的类别和频繁度等作为文本敏感信度计算因子，利用式（2）计算文本的敏感程度。

$$s'(t)=\frac{\sum_{i=0}^{n-1} s_{loc}(s_i)\times typ(s_i)}{n} \quad (2)$$

其中：$s'(t)$ 代表文本 t 的敏感程度；$s_{loc}(s_i)$ 表示敏感词 s_i 的位置敏感度；$typ(s_i)$ 表示敏感词 s_i 所对应类别敏感度；n 表示文本长度。

利用归一化方法将 $s(t)$ 的值映射到[0,1]区间，则文本 t 的敏感程度为

$$s(t)=e^{-\frac{1}{s'(t)}},\quad (i=1,2,\ldots,n) \quad (3)$$

3.5　文本敏感信息过滤算法

本文通过计算文本敏感程度，为网络平台处理文本提供参考。设定两个阈值 λ、ε（$\varepsilon<\lambda<1$），将上述公式计算出的文本敏感程度和设定阈值作比较，敏感程度高于 λ 的文本，为避免正向性文本被删除，其中含有较大数量的敏感词，例如一些反对敏感信息的文本中包含大量敏感词或政府机关用于抵制敏感词的文章，比如抵制贩毒的政府公文。进一步判断其文本的正向、反向性。若属于反向文本，直接从网络平台删除并追究其作者相应的责任；敏感程度介于 λ 与 ε 之间的文本，需要接收相关部分的审查；敏感程度低于 ε 的文本，则不需要作处理。文本自动过滤算法（automatic text filtering, ATF）具体的执行过程如算法 3 所示。

算法 3　ATF algorithm

输入：文体 t 。

输出：文本 t 的处理结果。

a）基于决策树输出叶子节点信息中的敏感词及其变形体集合 $S=\{s_0,s_1,\ldots,s_i,\ldots,s_{n-1}\}$，并输出其位置和类型，形成敏感词及其变形体位置信息集合 $Loc=\{l_0,l_1,\ldots,l_i,\ldots,l_{n-1}\}$。

b）计算每个敏感词或者敏感词变形体的位置敏感度，其中根据位置信息次数的累加，可得到频繁度信息 $s_{loc}(s_i)$。

c）基于敏感词类型表，计算每个词的类型敏感度 $typ(s_i)$。

d）取阈值 λ、ε（$\lambda>\varepsilon$）。

e）基于多因子模型，计算该文本 t 的敏感度 $s'(t)$。

f）对 $s'(t)$ 进行归一化，得到 $s(t)$。

g）当 $s(t)>\lambda$，则网络平台应判断文本正向、反向性。反向则删除文本 t 。当 $\varepsilon<s(t)<\lambda$，则 t 需要进行人工审查；当 $s(t)<\varepsilon$，则 t 不需要处理。

h）算法结束。

4　实验与分析

为了验证文本敏感程度计算方法的可行性，搭建了实验环境，选择了合适的数据进行实验，通过给定不同的实验条件，收集实验数据并对实验结果进行不同角度的分析。

4.1　实验环境

本实验在具有 2.4 GHz Inter(R)Core(TM)i7 处理器 8 GB 内存的机器上运行，操作系统为 Windows 10，编程工具为 Pycharm，编程语言为 Python。

4.2　数据集

为了评估面向中文敏感词变形体的识别方法的效果，本文从 CSDN(https://download.csdn.net)下载了含有疑似敏感词的 26 728 条新浪微博文本(包含科技、体育、金融、社会、娱乐等类型)作为测试数据集。首先对数据进行预处理，然后对数据集中的敏感词及其变形体进行人工的识别和分类，筛选出包含敏感词变形体的文本共 3 835 篇，其中共发现 554 个敏感词及变形体 1 288 个，涵盖了词的拼音、词的简称、词的拆分三种变形体情况，并将识别出的敏感词存入敏感词表中。数据集中所抽取的敏感词变形体的部分举例，如表 4 所示。

表 4　敏感词变形体的部分举例

Table 4　Examples of sensitive word deformed bodies

敏感词	词的拼音	词的简称	词的拆分
法轮功	falungong	flg	氵去车仑工力
兴奋剂	xingfenji	xfj	兴大田齐刂
贩卖毒品	fanmaidupin	贩毒 fmdp	贝反卖毒口口口
袭警	xijing	xj	龙衣敬言

4.3　实验分析

4.3.1 敏感词识别算法的对比分析

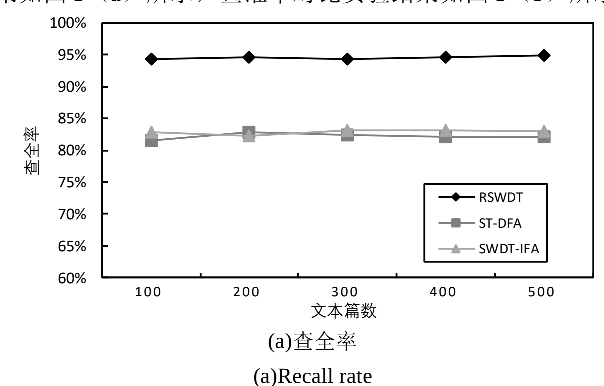
在实验 1 中，根据文本篇数和文本长度的变化，通过本文提出的 RSWDT 算法与基于确定有穷自动机的改进算法

（ST-DFA）、敏感信息决策树信息过滤算法（SWDT-IFA）两种算法在敏感词变形体查准率和查全率两个方面的比较来验证本算法的有效性。

a)文本篇数对敏感词及变形体识别的查全率和查准率的影响对比。

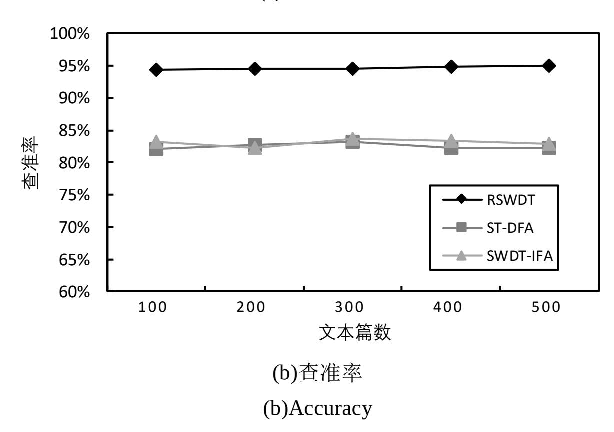
在本实验中选取含有敏感词变形体的文本 1 500 篇，将其随机分成五组进行测试：第一组为 100 篇，第二组为 200 篇，第三组为 300 篇，第四组为 400 篇，第五组为 500 篇。

RSWDT 与 ST-DFA 和 SWDT-IFA 的查全率对比实验结果如图 3（a）所示，查准率对比实验结果如图 3（b）所示。



(a)查全率

(a)Recall rate



(b)查准率

(b)Accuracy

图 3　查全率和查准率随文本篇数改变的变化趋势

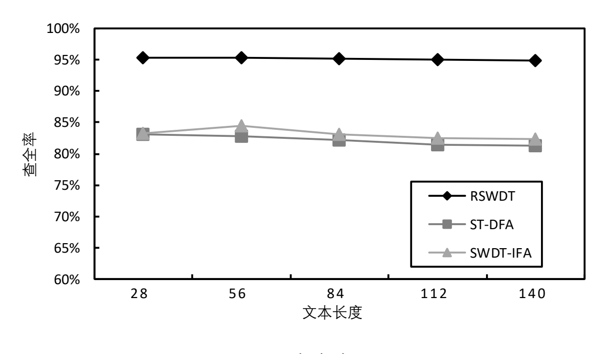
Fig. 3　Trend of recall rate and precision rate changes with number of texts

在文本篇数为 100 时，三种算法的查全率和查准率相比文本篇数为 500 时较低，可能是由于数据量太少，一个错误的识别就对实验结果有很大影响。RSWDT 算法随着数据的增长，查全率和查准率会慢慢趋于平稳。在文本篇数为 500 时，RSWDT 的查全率和查准率分别达到最高，达到了 95%。而 ST-DFA 算法随着文本篇数的增加，查准率较低且有下降趋势，SWDT-IFA 的查准率波动较大。总体看来 RSWDT 的查全率和查准率高于 ST-DFA 和 SWDT-IFA 算法。主要原因是本文提出的 RSWDT 算法不仅可以识别敏感词，而且可以有效识别敏感词变形体；ST-DFA、SWDT-IFA 算法虽然可以有效识别敏感词及简单的含有拼音的敏感词变形体，但对其余大部分敏感词变形体无处理能力。

b)文本长度对敏感词及变形体识别的查全率和查准率的影响对比。

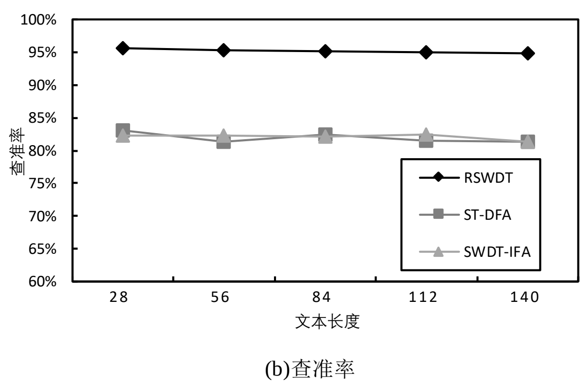
由于微博文本长度最大为 140 字，按照 28 个字为单位，将微博文本分为微文本（0~28 个字）、短文本（29~56 个字）、小文本（57~84 个字）、中文本（85~112 个字）及大文本（113~140 个字）共五类。在本实验中，从包含有敏感词变形体的文本中，对每类文本中各随机选取 500 篇进行实验。

RSWDT 与 ST-DFA、SWDT-IFA 的查全率对比实验结果如图 4（a）所示，查准率对比实验结果如图 4（b）所示。



(a) 查全率

(a)Recall rate



(b)查准率

(b)Accuracy

图 4　查全率和查准率随文本长度改变的变化趋势

Fig. 4　Trend of recall rate and precision rate changes with the length of text

在文本长度小于 28 个字时，三种算法的查全率和查准率非常高，而 RSWDT 算法随着数据的增长查全率和查准率会缓慢降低并趋于平稳。在文本长度介于 112~140 个字时，RSWDT 的查全率趋于稳定，达到了 95%。而 ST-DFA 和 SWDT-IFA 算法随着文本长度增加，查全率会波动并明显减少。RSWDT 的查准率稳定在 95%。而 ST-DFA 和 SWDT-IFA 算法随着文本字数的增加，查准率较低且下降趋势明显。总体看来 RSWDT 的查准率高于 ST-DFA 和 SWDT-IFA 算法。

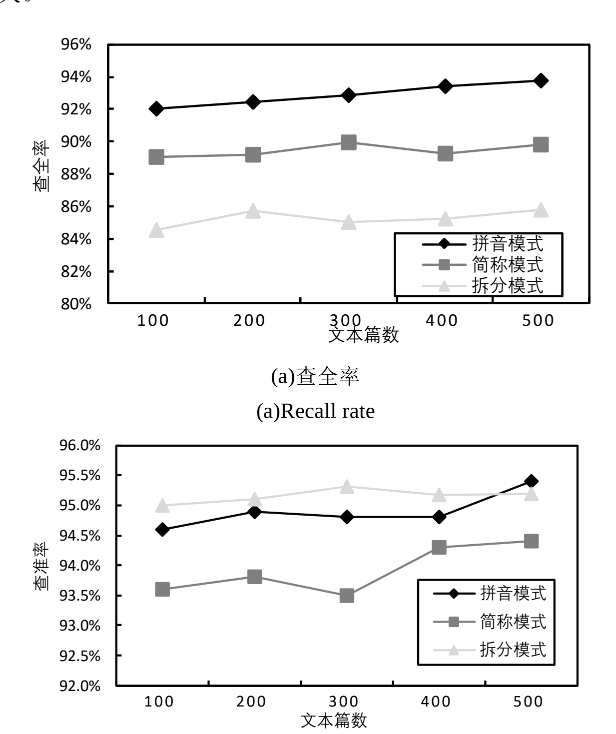
4.3.2 三种敏感词变形体识别有效性对比

在本实验中，选取含有敏感词变形体的文本 1 500 篇，将其随机分成五组进行测试：第一组为 100 篇，第二组为 200 篇，第三组为 300 篇，第四组为 400 篇，第五组为 500 篇。然后通过敏感词的三种变形体的查准率、查全率两个方面来验证本文提出的 RSWDT 算法的有效性。查全率对比实验结果如图 5（a）所示，查准率实验对比结果如图 5（b）所示。

在文本篇数为 100 时，三种变形体的查全率和查准率都比较低，可能是由于数据量太少，一个错误的识别就对实验结果有很大影响，所以随着数据的增长，查全率和查准率会慢慢趋于平稳。在文本篇数为 500 时，拼音模式的查全率最高，达到了 93%，拼音模式的查准率最高，达到了 94%。

从查全率来看，敏感词拼音模式的查全率高于简称模式，简称模式又高于拆分模式，主要原因可能是汉字的构造很复杂，人为拆分时对汉字的组成部件的分析不够全面，而简称中汉字组成方式过多。从查准率来看，敏感词拆分模式的查准率高于拼音模式，拼音模式又高于简称模式，主要原因可能是敏感词拆分模式虽然很复杂，但汉字组成结构比较固定，只要能识别出来基本就不会有错，敏感词简称模式的汉字和

拼音也比较固定，而拼音模式中易混拼音对结果的干扰比较大。



(a)查全率

(a)Recall rate

(b)查准率

(b)Accuracy

图 5　不同变形体的查全率和查准率变化趋势

Fig. 5　Trend of recall rate and precision rate of different deformed bodies

4.3.3 敏感度阈值设定对文本敏感信息过滤效果影响对比

为了验证文本敏感信息过滤的可行性，本文抽取 2 132 篇文本进行编号，然后随机分成 4 个样本，每个样本含有 533 篇，将每个样本分给 100 个人，要求这 100 个人对文本中的敏感词进行识别统计并对文本的敏感程度进行判断，判断结果分为三类，并分别用不同的符号进行标记，敏感度最高类可表示为 A，敏感度中等类可表示为 B，不需处理类可表示为 C。将 100 个人判断的每篇文本的敏感词个数取平均值，将 100 个人判断的文本类别取最多的类别作为最后的分类结果，如表 5 所示。

表 5　实验文本敏感程度表

Table 5　Table of sensitivity of experimental text

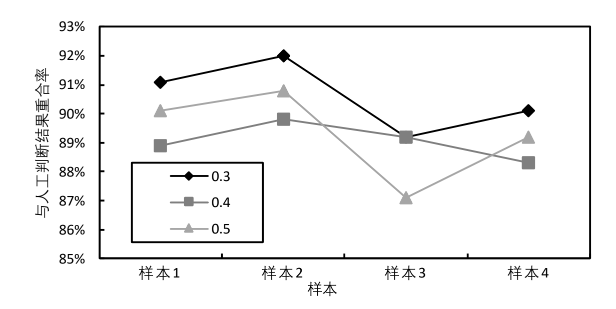
文本序号	敏感词个数	文本敏感程度类别
1	8	A
2	5	C
3	4	C
4	5	B
5	6	C
6	6	B
…	…	…
2132	7	A

分别对每个类别的敏感词个数取平均值，其中 A 类文本含有敏感词变形体平均个数为 7 个，B 类含有 4 个，C 类含有 2 个。从整理后的结果中可以看出，A 类文本的敏感词个数明显高于 B 类文本，B 类明显高于 C 类。由此可以得出人在一般情况下是根据敏感词个数来判断文本的敏感程度，文本中敏感词个数越多，则认为文本的敏感程度越高。

接下来用本文提出的敏感程度计算方法，对 2 132 篇文本进行敏感程度的计算，设阈值 λ、ε（$\varepsilon<\lambda<1$）。当实验结果大于 λ，该文本为 A 类；当实验结果小于 ε，该文本为 C 类；当实验结果介于 λ 与 ε 之间，该文本为 B 类。

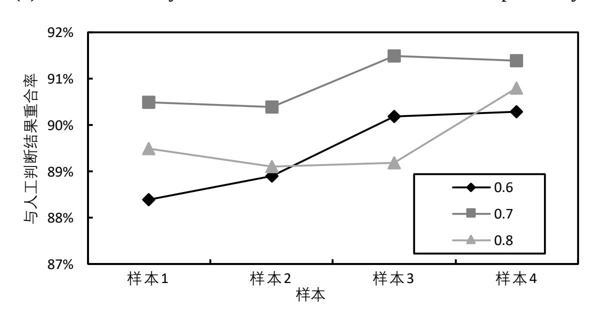
取高阈值 λ 为 0.8，低阈值 ε 分别取值为 0.3、0.4 和 0.5 进行三组实验，将结果与表 5 进行比较.当与人工判断结果相同时加 1，最后与每个样本中含有的文本篇数 533 求比值。实验结果如图 6（a）所示。

取低阈值 ε 为 0.3，高阈值 λ 分别取值为 0.6、0.7 和 0.8 进行三组实验，将结果与表 5 进行比较.当与人工判断结果相同时加 1，最后与每个样本中含有的文本篇数 533 求比值。实验结果如图 6（b）所示。



(a)敏感度低阈值分别为 0.3、0.4 和 0.5

(a) Low sensitivity thresholds were 0.3, 0.4 and 0.5, respectively



(b)敏感度高阈值分别为 0.6、0.7 和 0.8

(b) High sensitivity thresholds were 0.6, 0.7 and 0.8, respectively

图 6　不同阈值设置对文本敏感信息过滤效果的影响

Fig. 6　Influences of different threshold settings on filtering effect of text sensitive information

从图 6（a）中可以看出，当低阈值 ε 取 0.3 时与人工判断结果的重合度最高；从 (b)中可以看出，当高阈值 λ 取 0.7 时与人工判断结果的重合度最高。综上，当低阈值 ε 取 0.3、高阈值 λ 取 0.7 时，实验结果更接近人工识别的结果。

4.3.4 实验结论

实验结果表明，基于决策树的敏感词变形体识别算法（RSWDT）对中文敏感词变形体的识别有较高的准确率。通过对文本敏感程度的计算能够全面的体现文本的敏感性，不仅可以减少人工的工作量，也为含有敏感信息的文本处理提供了更直观可靠地依据，有效地过滤掉敏感信息，实现文本的自动过滤。

5　结束语

基于决策树的敏感词变形体算法能够有效的识别出词的拼音、词的简称和词的拆分三种敏感词变形体。基于多因子改进模型对文本进行敏感程度计算，从而达到文本自动过滤的效果。本文提出的算法有效地提高了敏感信息特别是敏感词变形体识别和过滤的准确率和效率。实验证明，其效果更接近于人工识别的结果。但是此类研究缺少对词与词之间的语义分析，需要人工对文本的正向与反向意思进行判断，当

敏感信息文本较多时，工作量会比较大，所以汉字之间的关系和语义是非常重要的，这也是本文下一步的工作。

参考文献：

[1] 刘梅彦，黄改娟. 面向信息内容安全的文本过滤模型研究 [J]. 中文信息学报，2017, 31 (2): 126-131. (Liu Meiyan, Huang Gaijuan. Research on harmful text filtering model based on semantic analysis [J]. Journal of Chinese Information Processing, 2017, 31 (2): 126-131.)

[2] 俞浩亮，王秋森，冯旭鹏，等. 基于特征加权的网络不良内容识别方法 [J]. 现代电子技术，2016, 39 (3): 76-79. (Yu Haoliang, Wang Qiusen, Feng Xupeng, *et al.* Feature weighting based identification method for network undesirable content [J]. Modern Electronics Technique, 2016, 39 (3): 76-79.)

[3] Wu O, Hu W. Web sensitive text filtering by combining semantics and statistics [C]// Proc of IEEE NLP-KE'05. [S. 1.] : IEEE Press, 2005: 215-259.

[4] 薛朋强，努尔布力，吾守尔•斯拉木. 基于网络文本信息的敏感信息过滤算法 [J]. 计算机工程与设计，2016, 37 (9): 2447-2452. (Xue Pengqiang, Nurbol, Wuxur • Islam. Sensitive information filtering algorithm based on text information network [J]. Computer Engineering and Design, 2016, 37 (9): 2447-2452.)

[5] 邓一贵，伍玉英. 基于文本内容的敏感词决策树信息过滤算法 [J]. 计算机工程，2014, 40 (9): 300-304. (Deng Yigui, Wu Yuying. Information filtering algorithm of text content-based sensitive words decision tree [J]. Computer Engineering, 2014, 40 (9): 300-304)

[6] Yoon T, Park S Y, Cho H G. A smart filtering system for newly coined profanities by using approximate string alignment [C]//Proc of IEEE International Conference on Computer & Information Technology. 2010: 643-650.

[7] Sood S 0, Antin J, Churchill E F. Using crowdsourcing to improve profanity detection [C]// Proc of AAAI Spring Symposium Series. 2012: 69-74.

[8] 付聪，余敦辉，张灵莉. 面向中文敏感词变形体的识别方法研究 [J]. 计算机应用研究，2019, 36 (4) . (Fu Cong, Yu Dunhui, Zhang Lingli. Study on the identification method for the change form of Chinese sensitive words [J]. Application Research of Computers, 2019, 36 (4))

[9] Chang J S, Teng Weilun. Mining atomic Chinese abbreviation pairs: a probabilistic model for single character word recovery [J]. Language Resources and Evaluation, 2007, 40 (3/4): 367–374.

[10] 中国文字改革委员会. 汉语拼音方案 [S]. 1967. (Chinese Character Reform Commission. Scheme of the Chinese phonetic alphabet [S]. 1967.)

[11] 殷志平. 构造缩略语的方法和原则 [J]. 语言教学与研究，1999 (2): 73-82. (Yin Zhiping. Methods and principles for the construction of abbreviations [J]. Language Teaching and Linguistic Studies, 1999 (2): 73-82.)

[12] 朱文轩. Blog 文本内容敏感信息的自动提取技术 [D]. 上海： 上海交通大学，2008. （Zhu Wenxuan. Automatic extraction technology of blog text content sensitive information [D]. Shanghai: Shanghai Jiaotong University, 2008.）

[13] Nghia L T, Huy A Q, Ngoc A N. Application of fuzzy-analytic hierarchy process algorithm and fuzzy load profile for load shedding in power systems [J]. International Journal of Electrical Power & Energy Systems, 2016, 77: 178-184.

[14] Lan S, Zhang H, Zhong R Y, *et al.* A customer satisfaction evaluation model for logistics services using fuzzy analytic hierarchy process [J]. Industrial Management & Data Systems, 2016, 116 (5): 1024-1042.